Перевод с английского языка: Охрименко К.С.

Крупномасштабная структура семантической сети: Статистический анализ и модели семантического роста

Марк Стейверс
Джошуа Б. Тененбаум
Отдел когнитивных наук Университета Калифорнии в Ирвине
Отдел мозга и когнитивных наук, Массачусетского технологического института
Источник: http://web.mit.edu/cocosci/Papers/03nSteyvers.pdf

Аннотация.

Мы представляем статистический анализ крупномасштабной структуры, состоящей из трех типов семантических сетей: ассоциация слова, WordNet и тезаурус Роджета. Мы продемонстрируем, что они имеют небольшую структуру, которая характеризуется разряженной связью, средней длиной пути между словами, и сильной локальной кластеризацией.

Кроме того, распределение числа соединений происходит по степенному закону, который позволяет определить размер (вес) соединения свободного образца, при этом большинство узлов, имеющих относительно небольшое число связей объединяются с помощью небольшого числа центров с большим количеством связей. Эти закономерности были также обнаружены в некоторых других сложных природных сетях, таких как World Wide Web, но они не согласуются с многими обычными моделями семантических сетей, основанными на иерархии наследования, произвольно структурированных сетей, или многомерных пространств.

Мы полагаем, что эти структуры отражают механизмы, посредством которых семантические сети растут. Мы опишем простую модель семантического роста, в которой каждое слово или понятие связано с существующей сетью и дифференцирует связи структуры существующего узла. Мировая статистика использования и применения данной модели еще достаточно невелика. Данная модель предлагает одну из возможных механических основ для изучения воздействия переменных (использования частот) на изменение производительности в задачах семантической обработки.

Ключевые слова: семантическая сеть, семантическое представление, рост сетевых моделей

1. Введение

Сетевые структуры обеспечивают интуитивно понятное и полезное представление для моделирования семантических знаний и выводов. В парадигме семантических моделей сети, мы можем выделить три типа вопросов.

Первый тип вопросов касается структуры и знаний: в какой степени организация семантического знания человека может быть объяснена с точки зрения общих структурных принципов, которые характеризуют связи семантической сети?

Второй тип вопрос касается процессов и производительности: в какой степени представление человека о семантических задачах обработки могут быть объяснены в рамках общих операционных процессов семантических сетей?

Третий тип вопросов касается взаимодействия структуры и процессов: в какой степени процессы семантического поиска и извлечения информации используют общие структурные особенности семантической сети, и в какой степени эти структурные особенности отражают общие процессы семантического роста или развития?

Самые ранние работы над семантическими сетями отвергали решение данных вопросов. Коллинз и Кьюллиан (1969) предположили, что понятия представлены в виде узлов древовидной иерархии, со связями определяемыми классами включения отношений (рис. 1).

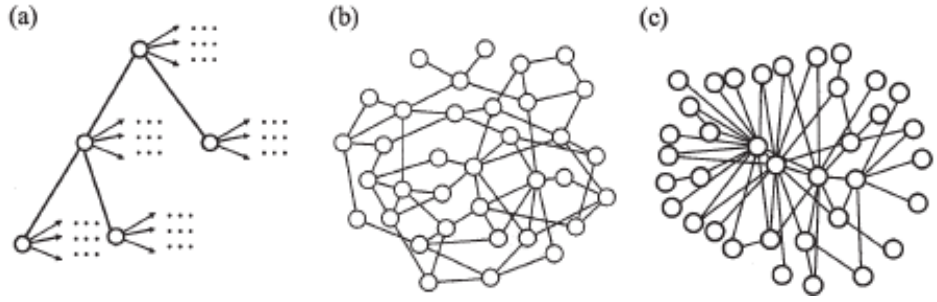

Рисунок 1 – Предлагаемые крупномасштабные структуры для семантических сетей: (а) – древовидная иерархия; (б) – произвольный неструктурированный граф; (с) – граф со свободным масштабом.

Дополнительные узлы характерных признаков или предикатов связанны с наиболее общим уровнем иерархии, к которым они применяются. Особенность древовидной иерархии – экономичность системы по умолчанию для представления знаний о категориях, но это накладывает сильные ограничения на возможности расширения предикатов, по содержанию, по видам возможных знаний (Кэйл, 1979; Соммерс, 1971).

Коллинз и Кьюллиан предложили алгоритмы для эффективного поиска этих иерархий наследования для получения или проверки фактов. Они

показали, что время реакции человека в качестве субъекта часто являтся достаточно качественным предсказанием этой модели. Однако, несмотря на изящество этой картины, она имеет жесткие ограничения, как общая модель семантической структуры. Наследование иерархии явно подходит только для определенных таксономически организованных понятий, таких как классы животных или другие природные виды. Даже в тех идеальных случаях, строгой структуры наследования, кажется, не применяются за исключением наиболее типичные элементы иерархии (Кэри, 1985; Коллинз и Къюллиан, 1969; Рипс, Шобен, и Смит, 1973; Сломан, 1998).

Последующие работы по семантическим сетям перешли от поиска общих структурных принципов знания организации к выяснению механизмов семантической обработки произвольно структурированных сетей.

Сетевые модели Коллинза и Лофтуса (1975), например, не отличаются от любой крупномасштабной структуры – древовидной иерархии. Связи крупномасштабных сетевых моделей по существу не структурированы с каждым словом или концептом соответствующего узла ,а также связями между любыми двумя узлами, которые непосредственно связаны каким-то образом (рис. 1б).

Количественные модели общей ассоциативной сети, часто оснащены процессом распространения активации, который используется для прогнозирования производительности в ряде экспериментальных задач поиска памяти и объясняет различные явления заполнения и интерференции (Андерсон, 2000; Колинс и Лофтус, 1975; Диз, 1965 Нельсон, Мак Кинни, Джи и Янцзура, 1998).

В настоящее время в ходе исследований по данному вопросу еще не пришли к единому выводу, по крайней мере в некоторых из процессов, связанных с формированием и поиском семантической памяти (Anderson, 2000). С другой стороны, существует соглашение об общих принципах управления крупномасштабной структурой семантической памяти, и есть описание, как эта структура взаимодействует с процессами поиска памяти или приобретения знаний. Типичные иллюстрации учебника семантической памяти до сих пор отображают по существу произвольные сети, такие как на рис. 1b, без отличительных крупномасштабных структур.

Содержание семантической сети неотделимо от структуры содержания концепта, по крайней мере в месте соединения с другими понятиями. Таким образом, без общей структуры принципов, парадигма семантической сети малоэффективна или не дает общего понимания в характере семантики.

В этой статье мы утверждаем, что на самом деле существуют достаточно убедительные общие принципы, регулирующие структуру сети представления семантики естественного языка, и что эти структурные принципы имеют потенциально серьезные последствия для процессов семантического роста и памяти поиска.

Мы подчеркивали, с самого начала, что эти принципы не предназначены для обеспечения подлинной теории семантики, и мы не

верим, что сеть четкие отношения, точно отражающие все наиболее важные и глубокие аспекты семантической структуры. Мы предполагаем, что семантические сети будут играть определенную роль в любом процессе определения значения слова. Наша цель – изучить некоторые общие структурные свойства семантических сетей, которые могут в конечном итоге стать составной частью основы для любой семантической теории.

Принципы, которые мы предлагаем, не основаны на любых фиксированных структурных рассуждениях, таких как древовидные структуры (иерархии) Коллинза и Къюллиана (1969). Скорее, они основаны на статистических закономерностях, которые мы обнаружили с помощью анализа теории графов, в описанных выше работах по семантическим сетям.

Мы рассматриваем распределение ряда статистик, которые рассчитываются на узлы, пары узлов, или тройки узлов семантической сети: количество связей с словом, длина кратчайшего пути между двумя словами, и процент соседей узла, которые сами выступают соседями.

Эти статистические принципы семантической структуры сети носят довольно общий характер. Они справедливы и для семантического представления сети, построенной другими способами (Нельсон, Мак Эвой, и Шрайбер, 1999) или рассматриваемого анализа лингвистов (Миллер, 1995; Роджет, 1911). В то же время, эти закономерности не могут использоваться для многих распространенных моделей семантической структуры, в том числе и иерархической или произвольно (неструктурированных) связанных сетей (рис. 1a и 1b), а также моделей высокой размерности многомерного пространства, такие как латентный семантический анализ (LSA; Ландауэра и Дюмэ, 1997).

Таким образом эти принципы могут стать основой для новых подходов к моделированию, а также для расширения или пересмотра существующих моделей. В конце концов, они могут помочь определить, какие классы моделей, наиболее точно отражают структуру семантики естественного языка.

В нашей модели сеть приобретает новые понятия с течением времени и связывает каждый новый концепт подмножества понятий в рамках существующего соседства, с вероятностью выбора особенности окрестности пропорционально его размеру. Этот процесс роста можно рассматривать как своего рода семантическую дифференциацию, в которой новые понятия соответствуют более конкретной вариации существующих концептов, и очень сложные понятия (со множеством связей), более вероятно, должны быть дифференцированы, чем простые рис.1C